MedCrave
Step into the World of Research

MOJ Sports Medicine

Research Article

Open Access

CrossMark

Cluster methods in function better selection

Summary

Cluster analysis methods, also known as taxonomic methods, are intended for grouping objects and subjects according to certain characteristics, attributes and properties. Cluster analysis looks at relevant objects and attributes, classifying them into two or more independent groups. Cluster analysis supplemented with discriminant analysis is used in confirmatory and fundamental research. In numerous statistical-methodological procedures, these methods are applied when setting up and testing various hypotheses. Grouping methods are particularly useful in the process of different selections with the aim of forming coherent groups, which may or may not necessarily be statistically different. There are several models of clustering (grouping), always with one goal, which is greater proximity (similarity) of an entity belonging to a group compared to an entity belonging to another group. Two basic grouping models are recognizable, Hierarchical and Non-Hierarchical. Both models have the same goal, which is the formation of several independent homogeneous groups from one common group of entities. The hierarchical approach does not define the number of clusters in advance (a priori), in contrast to the Non-Hierarchical Model which defines in advance number of clusters. The grouping model is chosen depending on the specific problem and the set goal of grouping. In the process, several different models are often applied, and then one is chosen as in this research. It is important to point out that the theoretical number of clusters (groups) is often not realistically applicable in practice. Using the example of this research, it was proven that the first grouping was not a good solution. Through the subsequent, second and third iteration, as well as the application of additional discriminative methods, three optimal clusters were determined in the population of girls and boys. Satisfactory optimal grouping was obtained on the basis of gender criteria and achieved results on psycho-motor tests.

Keywords: clusters, selection, students, psychomotor skills

Volume 7 Issue 2 - 2024

Mihajlo Mijanović
University of Podgorica, Montenegro

Correspondence: Mihajlo Mijanović, Faculty of Sports and Physical Education, University of Podgorica, Montenegro, Tel 38268650974, Email mihajlo.mijanovi@gmail.com

Received: April 01, 2024 | **Published:** May 01, 2024

Introduction

Cluster analysis1 belongs to statistical-mathematical methods whose main goal is to group objects based on the relevant characteristics of those objects. Cluster analysis is also known as taxonomic analysis, Q-analysis, Classification analysis and Numerical taxonomy.[1] Although the names are different, the common goal of these methods remains the same, which is the classification of objects or subjects based on relevant characteristics (characteristics). Grouping methods are widely used in various fields and scientific disciplines such as: anthropology, philosophy, sociology, psychology, medicine, education, healthcare, sports, art, architecture, demography, criminology, economics, business, trade, traffic, innovative technology, etc. Cluster analysis can be used when formulating hypotheses regarding the structure of objects, where hypothetical clusters are compared with obtained clusters. It is a confirmatory technique. Cluster analysis is also used to simplify data, by analyzing groups of similar objects instead of individual objects. The resulting structure finds relationships that cannot be identified otherwise. Cluster analysis can also be applied in fundamental research. The possibilities of applying cluster analysis are: large, which is why it is used in solving various professional and scientific problems. In addition to the many advantages of practical and scientific applicability, some critics believe that cluster analysis is not based on statistical inference, which is not scientifically based and empirically confirmed, on the contrary. The variables used in the analysis are particularly sensitive to changes in the observed characteristics. The choice of variable observations is non-standard (a personal matter), so there is no single method for data analysis using cluster analysis. The basic methodology of cluster analysis is the grouping of objects (subjects) into closely related clusters (groups), so that objects located in the same cluster are more similar to each other compared to objects located in other clusters. The idea of this analysis is to maximize the homogeneity of objects within a cluster, while at the same time maximizing the heterogeneity between other, other clusters. Cluster analysis observes selected relevant variable characteristics, but does not separate them into dependent and independent. Based on the selected variable observations, the procedure of grouping into separate clusters is carried out. The number of objects and features is virtual. Objects and features are mutually independent. An example of objects that are a frequent object of observation and study can be all living and non-living beings, states, regions, cities, etc. Entities are individual instances of a set that have specific characteristics. So the basic carrier of information is an entity. If the basic set consists of people, students, the entity is one student, and the feature or characteristics of student observation can be: class, success, place of residence, social status, number of family members, sports results, somatic status, height, weight, health status, etc. In any case, the number of features is infinite, and only relevant features for a specific problem are taken into consideration. Therefore, cluster identification is performed on the basis of pre-selected relevant characteristics (characteristics) of specific objects (entities). The goal of Cluster Analysis is to divide the basic set (population) into a certain number of groups or clusters, so that all the objects-entities of one cluster are closer and more similar in terms of the observed feature compared to any other cluster. The methodological grouping procedure is carried out with the help of a classification algorithm that enables classification.[2] The input data contains the population matrix [Y] on the basis of which the selection is made into (k) different groups, i.e. clusters. As a rule, the population matrix is rectangular, where the rows (species) contain objects, and the columns (columns) contain relevant characteristics (observations).

Research design in cluster analysis

Before starting the process of dividing objects or subjects based on some characteristics (characteristics), it is necessary to define the objectives of the research and carry out the selection of objects and relevant characteristics of the objects. Statistical terminology refers to the selection of objects that can be people or any objects of observation: cities, regions, settlements, buildings, schools, health facilities, institutions, sports clubs, as well as the belonging subjects of those areas. The application of a certain method of cluster analysis (two-stage cluster analysis and k-means cluster analysis) depends on the arrangement of objects in the database. Different arrangement of data gives different results, so it is necessary to arrange the data by random selection method. The smaller the database, the greater the problem of order of objects, even if they are arranged in a random way. Therefore, when conducting the analysis, it is proposed to redistribute the objects. We often encounter categorical variables in cluster analysis. If there is at least one variable, i.e. categorical variable, then a two-stage cluster analysis is recommended.[3] It is especially important to take into account whether there are non-standard observations and whether they should be excluded, as well as whether standardization of variables is required. Non-standard observations can be observations that are not representative of the population, but that are representative of the specific sample and research problem. For the above reasons, a preliminary data review and preliminary analysis is necessary.

Assumptions of cluster analysis

Cluster analysis is not a multivariate technique based on statistical locking, where unknown parameters of the population are evaluated through sample statistics. Cluster analysis is a method for quantifying the structural characteristics of a set of objects based on specific observations. This method has mathematical properties, but not a statistical basis.[4] The conditions of normality and linearity, which are very important in many statistical techniques, have very little importance in cluster analysis. That is why it is important for cluster analysis to determine whether a representative sample is used and whether there is multicollinearity of variables. Only in rare cases can cluster analysis be performed on the total data. Mostly, cluster analysis is carried out on samples or subsamples, taking care that the samples are representative of the entire population. This is a very important condition because only on the basis of representative samples we can generalize conclusions to the entire population.[5]

Multicollinearity represents the degree of association of an independent variable with other independent variables. Increasing multicollinearity reduces or excludes the possibility of defining the influence of one variable in a particular analysis. In cluster analysis, the effect of multicollinearity is of a completely different form because multicollinearity represents a form of indirect weighting. So the multicollinear weighting process is not obvious, but it affects the analysis. That is why it is necessary to perform an analysis of the significance of multicollinearity. If there is multicollinearity in that case, it is necessary to reduce the variables to one variable for each set of correlated variables, or to use a distance measure that compensates for the correlation, i.e. it is necessary to apply the "Mahalanobis distance". When there are more than two variables, it is necessary to determine the calculated multidimensional position of each observation in relation to some common point, for this we use the Mahalanobis $\boldsymbol{D^2}$ measure, which measures the distance of each observation in the multidimensional space from the central or average center of the observed variables. A high value of $\boldsymbol{D^2}$ indicates non-standard observations.[6] To define this measure, we can assume that we have a random variable y that has a normal distribution with an arithmetic mean of zero and a variance or standard deviation of one. If there are two variables, yi and yj, to compare the distance, it is necessary to take into account the variance of the random variable. Then the squared distance between yi and yj is defined as:

$$D^2_{ij} = \left(y_i - y_j\right)^2 / \sigma^2$$

where σ^2 is the variance of the population. The assumption of multicollinearity can be tested via a bivariate correlation matrix for quantitative variables, to determine whether the correlation coefficients are statistically significantly different from zero.[7] Objects or observations are assumed to be mutually independent. The values of object k should not affect the values of object 1+k, that is, there is no autocorrelation of objects.

Closeness measures (similarities)

The concept of closeness is a basic concept of cluster analysis. Closeness between objects is an empirical measure of correspondence between objects that should be grouped into clusters. The analysis process takes place by calculating a closeness measure for all pairs of objects, where the closeness is based on the profile of each observation for the characteristics (variables) chosen by the researcher. In this way, each object is compared to any other object via a measure of closeness. The cluster analysis procedure continues to group similar objects into clusters. Closeness can be expressed through similarities or differences. If a closeness measure shows similarity, the value of the measure increases when two objects are more similar. Conversely, if the measure of closeness indicates dissimilarity, the value of the measure decreases when two objects are more similar. For two objects yi and yj in p-dimensional space, the similarity measure satisfies the following letters:

$\left(0 \le s_{ij}\right)$, for all objects y_i and y_j

$\left(s_{ij} = 1\right)$, if and only if y_i and y_j

$\left(s_{ij} = s_{ji}\right)$

Conditions one and two indicate that the measure is always positive and equal to unity if objects i and j are identical, while the third condition indicates that the measure is symmetrical. If there is a measure of similarity that satisfies the stated conditions, it is always possible to observe a measure of dissimilarity, i.e. $d_{ij} = 1 - s_{ij}$ on the contrary, if a measure of dissimilarity dij is known, it is possible to construct a measure of similarity as $s_{ij} = 1/(1 + d_{ij})$ In this way, it is possible to obtain the measure sij depending on the measure dij. Therefore, it is possible to use the measure of similarity as well as the measure of dissimilarity in cluster analysis. A well-known measure of similarity is the **"Pearson correlation coefficient"** between object y_i and y_j, $\left(i, j = 1,2,3,...n\right)$, defined by the formula:

$$\frac{\left[\sum_{s=1}^{p}\left(y_{is} - \overline{y}_{i.}\right)\left(y_{js} - \overline{y}_{j.}\right)\right]}{\left[\sum_{i=1}^{n}\left(y_{is} - \overline{y}_{i.}\right)^2 \sum_{i=1}^{n}\left(y_{js} - \overline{y}_{j.}\right)^2\right]}$$

Similarity measures of binary variables are important in cluster analysis. In order to construct measures of similarity of binary data, a contingency table of binary variables is used. The question arises how to weight pairs of the same and pairs of different codes of binary variables, because the pair (1–1) can be more significant than the pair (0–0). The first pair indicates the presence of the feature, while the second pair indicates the absence of the feature. It is also possible that the pairs (0–0) are not considered in the analysis at all. To allow for the different weighting of pairs of same and pairs of different binary

variables, as well as the treatment of (0–0) pairs, there are several different distances.[8]

Measures of distance (difference)

Let y_i and y_j represent two objects in space with p- variables. The diversity measure satisfies the following conditions:

$\left(d_{ij} \geq 0\right)$, d for all objects yi and yj

$\left(d_{ij} = 0\right)$, if and only if $y_i = y_j$

$\left(d_{ij} = d_{ji}\right)$

The first condition indicates that the measure is never negative. The second condition indicates that the measure is equal to zero when the objects are equal to each other, that is, the objects are equal only when dij=0 and in no other situation. The third condition indicates that the measure is symmetric, so that the dissimilarity measure comparing y_i with y_j is the same as the dissimilarity measure comparing objects y_j and y_i . A dissimilarity measure that satisfies these conditions is a semimetric measure. For numerical variables measured as a minimum on an interval scale, the most common measure of dissimilarity is the Euclidean distance between two objects. If we have a matrix [***Y***] of dimensions ($n \times p$) and a vector of dimensions y_i ($1 \times p$), the squared Euclidean distance between the two types y_i and y_j is defined as This process eliminates the dependence of the analysis on the measurement scale. But very often it causes the distances within the clusters to be greater than the distances between the clusters, and in this way the clusters overlap. The Euclidean distance[9] is a special case of the Minkowski metric, where the dissimilarity measures can be represented as: $d^2_{ij}=(y_i-y_j)(y_i-y_j)$. The data matrix [***D***]=d_{ij} dimensions (nxn) is called the Euclidean distance matrix. It is possible that different measurement units of the variables influence that a certain variable dominates in the quantification of the distance. The Euclidean distance matrix is most effective when the variables are expressed on the same measurement scale. If the variables are different and use different measurement scales, it is possible to calculate the weighting of the squared differences. A special case of the ***Euclidean distance*** is the *Minkowski, Canberra and Chekanowski distances*[10]. These measures are used when the data are asymmetric and/or when there are non-standard observations. The displayed measures are used in a situation where the variables are quantitative. For the categorical type of data measured on nominal or ordinal scales, the situation is more complex. In the simple case, it is assumed that each row yi of matrix [***Y***] contains only binary data. In this case, the squared Euclidean distance counts pairs that contain different binary values, (1–0) or (0–1), while treating pairs that have the same binary values, (1–1) or (0–0) equally. When a variable is coded with 0 or 1, it indicates the absence or presence of a particular characteristic. There are several methods and algorithms of cluster analysis. Two approaches known as *Hierarchical* and *Non-Hierarchical* cluster analysis dominate, i.e. hierarchical and non-hierarchical methods.

Hierarchical method

First, the distances of all units are calculated from each other, and then the groups are formed through merging or splitting techniques. The merging technique (*agglomerative, hierarchical method*) starts from the fact that each unit is alone in a group of one member. Close groups are gradually merged until eventually all units are found in one group. With the separation technique, the order is reversed, where two groups are created from one group, then the next two from those two, and so on until each unit of observation is separate. It is the so-called divisional hierarchical method, which is less often applied than the agglomerative one.[11]In a non-hierarchical approach, observation units can move from one group to another in different phases of analysis. There are many variations in the application of this technique, but the point is to first find a clustering point around which the units are located, in a more or less arbitrary way, and then calculate new clustering points based on the average value of the units. The observation unit is then moved from one group to another if it is closer to the newly calculated grouping point. The process takes place iteratively, until stability is reached for a predetermined number of groups. As there are several models of cluster analysis, it is good to implement several clustering methods in order to decide on the one that provides optimal solutions in a specific problem. Hierarchical methods group objects into the closest cluster at an early stage of clustering, but the same object cannot be regrouped into another cluster at a later stage, although this is a better solution. Such regrouping is possible with the non-hierarchical method. The advantage of the hierarchical method is that the number of clusters is not known a priori. That is why this technique is known as exploratory, while non-hierarchical methods are known as confirmatory. In the framework of hierarchical cluster analysis, agglomeration hierarchical methods and connection methods were observed.[12] Agglomeration hierarchical methods use the elements of the proximity matrix to generate a tree diagram or so-called dendrogram, where objects are combined into clusters, starting from the most similar objects to the least similar objects to finally obtain a single cluster. The cluster formation process begins with the formation of the distance matrix [$\boldsymbol{D}_{nxn}$] = (d_{ij}), and is carried out through the following steps.[13]

1. The process starts with n clusters, where each cluster contains one object.
2. The dissimilarity matrix [***D***] is sought, on the basis of which the most similar pair of elements is determined. The most similar pair is represented by the group (d_{ij}), whose objects i and j are chosen as the most similar.
3. The most similar pair is represented by a new cluster according to a certain criterion. In this way, the number of clusters is reduced by 1, by deleting rows and columns for objects i and j. The measures of dissimilarity between the formed cluster (ij) and all other clusters are calculated, using a certain criterion, and a row and column are added to the new dissimilarity matrix.
4. Steps 2 and 3 are repeated ($n-1$) times, until all objects form a single cluster. In each step, the merged clusters and the dissimilarity value are identified on the basis of which the clusters are merged. By changing the criteria in the third step, we obtain several hierarchical clustering methods that define the closeness between clusters.

Single or single connection (Nearest neighbor method)

To implement the nearest neighbor method, objects are combined into clusters using the least dissimilarity between clusters. If any element of cluster ***R***, $i \in \boldsymbol{R}$, and s is any element of cluster ***S,*** $j \in \boldsymbol{S}$, the distance between ***R*** and ***S*** is calculated:

$$d\left(\mathbf{R}\right)(\mathbf{S}) = min.[d_{ij}(i \in R, j \in S)]$$

With each step of the process, a dendrogram can be created, which is a graphic representation of the distances at which objects are connected. The branches of the tree represent clusters or objects. Branches are connected to nodes whose position on the similarity (or distance) axis indicates the level of connectivity.[14]

Complete connection (Farthest neighbor method)

The clustering procedure for complete linkage is the same as for individual linkage, except that at each stage, the distance (similarity) between two clusters is determined by the distance (similarity) between the two elements from each cluster, which are farthest from each other. The agglomeration algorithm starts by searching for the smallest distance in the distance matrix $[\boldsymbol{D}]=(d_{ij})$ and continues by merging the corresponding objects, i and j, to obtain a cluster (*ij*). In this case, if $i \in \boldsymbol{R}$ and $j \in \boldsymbol{S}$, where ***R*** and ***S*** are two clusters, the distance between clusters ***R*** and ***S*** is calculated[15]

$$d(\mathbf{R})(\mathbf{S}) = max.[d_{ij}(i \in R, j \in S)]$$

Average connection (Average Distance)

The input to the average linkage algorithm can be distances or similarities (distance or closeness). The process starts with the distance matrix $\boldsymbol{[D]}=d_{ij}$ to find the most similar objects, *i* and *j*. These objects are merged into a cluster (*ij*). The differences between this cluster and another cluster are determined by the cluster algorithm.[16] Instead of using the minimum or maximum as a measure, the distance between two clusters is calculated via the average dissimilarity value for each cluster:

$$[D]_{(R)(S)} = \left[\Sigma_i \Sigma_j d_{ij}\right]\left(n_R n_S\right)^{-1}$$

where $i \in \boldsymbol{R}$ and $j \in \boldsymbol{S}$, $n\boldsymbol{R}$ and $n\boldsymbol{S}$ represent the number of objects in each cluster.

Centroid method

In average linkage methods, the distance between two clusters is defined as the average value of the dissimilarity measure. If cluster ***R*** is assumed to contain n_g elements and cluster ***S*** to contain n_S elements, then the centroids for the clusters containing the two objects form the ***Squared Euclidean distance*** between the two clusters.[17]

$$d_{ij}^2 = \left[\overline{y}_i - \overline{y}_j\right]^2, [D] = (d_{ij})$$

The centroid agglomeration method starts with the distance matrix [D]=(d_{ij})Then the two most similar clusters are joined via the weighted average for the two clusters. If we mark the new cluster with C, then we calculate the centroid of the cluster using the formula:

$$\overline{y}_i = \left[n_R \overline{y}_i + n_S \overline{y}_j\right] / \left[n_R + n_S\right]$$

The centroid method is also known as the median method, if the unweighted centroid average is used, y¯_i=(y¯_i+y¯_j)/2. The median method is better when n_R is different from n_S, $(n_R \neq n_S)$.

Ward's method

Ward's method tends to find compact clusters of approximately the same size, noting that the cluster solution may be influenced by non-standard observations. This method uses hierarchical clustering procedures that minimize information loss when merging two groups. The loss of information means an increase in the value of the criterion - sum of squared error ***(SKG)***. For one cluster, ***R***, the ***RSKG*** sum of squared error is the sum of squared distances of each cluster object from the center (centroid) of the cluster. If the number of clusters is *k*, then:

$$kSKG = SKG1 + SKG2 + SKG3 + \ldots + SKGk.$$

At each step of the analysis, every possible union of clusters is taken into account, and two clusters whose combination results in the smallest increase in ***SKG*** (*or minimum loss of information*) are merged. Initially, each cluster has only one object, so if there are a total of n objects, then ***SKGk***=0, *k*=1,2,3…*n*, so ***SKG***=0. Conversely, if all clusters are in one group with n objects, then the ***SKG*** value is calculated by the formula:

$$\boldsymbol{SKG} = \sum_{i=1}^{n}\left(y_i - \overline{y}\right)'\left(y_i - \overline{y}\right) = \sum_{i=1}^{n}\left[y_i - \overline{y}\right]\wedge$$

where the multivariate measure yi is the i-th object, while is ¯y he mean of all objects.[18]

Non-hierarchical cluster analysis

In the hierarchical cluster method, the number of clusters is not known in advance. The process starts with a distance matrix, and once an object is grouped into a cluster, it is not reallocated. These methods can be used for both object grouping and variable grouping. Non-hierarchical cluster methods are used only to group objects. The process starts with the original data matrix [***Y***]. The cluster number k must be known in advance, as well as the cluster centroids or cluster cores, so that objects can be regrouped using a certain criterion, as well as marking the end of reallocation using certain rules to stop further regrouping of objects.

The most popular non-hierarchical method is ***Centroid Cluster Analysis.*** To initiate the non-hierarchical centroid cluster method, one must first determine the number of centroids (clusters) k. It is important to point out that the initial centers (points) of the clusters are random. Objects belonging to a cluster can be transferred to another cluster through an iterative process, it is about grouping and regrouping objects from cluster to cluster. The basic steps are[19]

1. Selection of k centers (cores) of clusters.
2. Grouping each object to the nearest centroid and again centroid calculation.
3. Step two is repeated until all observations are clustered or until the differences in cluster centroids become small enough.

The choice of cluster center can be done in two ways. The first is when the clusters are not determined by the researcher, which is the case when the data were analyzed using some other multivariate method. The most common example is using a hierarchical cluster algorithm to get the number of clusters and then generating the cluster center. It is considered that if the number of clusters is known, information about the basic characteristics of the cluster is also available. Another way of obtaining the center of the cluster is to generate it from the observations of the sample, in a systematic way or simply by random selection. The choice of cluster center is very important because different cluster solutions are obtained for different cluster centers. By applying cluster analysis for object grouping, hierarchical and non-hierarchical cluster methods can be combined. In the first step, a hierarchical procedure is used to identify the centers and the number of clusters, which would be input in a non-hierarchical procedure to obtain better analysis results. Centroid cluster analysis implies and assumes a large sample.

Two-level cluster analysis

Two-level cluster analysis is a method used when dealing with a large database, because hierarchical and non-hierarchical cluster methods do not have the same efficiency for large databases. This analysis is used for both categorical and numerical variables, and finds application in the analysis of categorical variables with three or more modalities. Two-stage cluster analysis is a method that requires only one pass through the database. The process consists of two steps. First, the object is initially grouped into smaller sub-clusters, and then

these sub-clusters are treated as special objects that are grouped using the hierarchical cluster method. It is possible for the two-stage cluster analysis algorithm to determine the number of clusters, but also for the number of clusters to be previously determined.

If a categorical (qualitative) variable is found in the two-stage cluster analysis, then the distance is determined by calculating the natural logarithm to obtain the Credibility Function. Objects with the highest values of this measure form a cluster.[20] If the variables are numerical (quantitative), then the Euclidean distance is used. Objects that have the smallest distance form a cluster. Euclidean distance is compatible for categorical qualitative and numerical quantitative variables. Calculation of distance measures is required in the first initial grouping, and in the second final grouping step. Distance calculation based on the natural logarithm, i.e. The credibility function is represented by a distance that is based on a probability. The distance between two clusters is related to the decreasing value of the natural logarithm, so the objects are merged into one cluster. It is desirable that Credibility Functions have a normal distribution for quantitative variables, or a multinormal distribution for categorical variables.

Two-level cluster analysis gives good results, even when the assumption of normality is not fulfilled. Also, this cluster analysis assumes that the sample is large. It is also assumed that the variables are mutually independent.

Determining the number of clusters

A very important issue in hierarchical and non-hierarchical analysis is determining the number of clusters that are most representative of the data structure. In the hierarchical analysis, a set of possible cluster solutions is created, but it is necessary to select one or several solutions that would most adequately represent the structure of the objects. The same decision is made in non-hierarchical analysis, where the best solution is chosen between two or more offered cluster solutions. There is no standard objective selection procedure for choosing the best solution. Many criteria have been developed that use complex approaches and are characteristic of different software programs. One group of these criteria is measures of change in heterogeneity. These measures follow changes and are calculated throughout the course of cluster merging, and are used when we have a significant increase in heterogeneity, because previously merged clusters are considered to be the best solution. This is also logical, because when clusters that significantly increase heterogeneity are merged, it is obvious that the previous solution was better. In Regression Analysis, the coefficient of determination $\boldsymbol{R^2}$ shows the percentage of variability explained by regression in relation to the total variability, that is, the degree to which the variations of the dependent variable are explained by the variations of the independent variable. In the Analysis of Variance, the $\boldsymbol{R^2}$ coefficient is defined as the ratio between the sum of the squares of the groups and the total sum of the squares and is a measure of the total variation of the dependent variable that is contained or explained through the arithmetic means of the groups. So in cluster analysis we can construct $\boldsymbol{R^2}$ and calculate it every time the number of clusters changes. So for n clusters, the total sum of squares $\boldsymbol{T}$ is[21]

$$T = \sum_{i=1}^{n} \left[y_i - y_j \right] \wedge$$

The sum of squares between the SKG clusters is:

$$SKG = \sum_{i=1}^{k} \left[y_i - y_k \right] \wedge$$

The R2 coefficient for k clusters is:

$$R_k^2 = \left[T - \sum\nolimits_K SKG_K \right] T^{-1}$$

For n clusters, it is valid that SKGk=0 , so that R^2=1 so that the number of clusters decreases from n to 1. With this procedure, the clusters should become more and more different. A large decrease in $\boldsymbol{R^2}$ should indicate a specific difference between clusters. When merging clusters $\boldsymbol{R}$ and $\boldsymbol{S}$, the *Semi-Partial Index of Differences* can be calculated:

$$SR^2 ie. SR^2 = R^2{}_k - R^2{}_{k-1}.$$

The $\boldsymbol{SR^2}$ statistic compares the ratio $\boldsymbol{SKG_t - (SKG_r - SKG_s)}$, where $\boldsymbol{C_R}$ and $\boldsymbol{C_S}$ clusters are combined to form $\boldsymbol{C_T}$ in the total sum of squares.

$$T = \sum_{i=1}^{n} \left[y_i - \overline{y} \right] \wedge$$

The greater the increase, the greater the *"loss of homogeneity",* that is, the clusters are more separated. In the analysis, we use several statistical tests to obtain the degree of heterogeneity. Each new cluster is created by merging two previous clusters. The most frequently used heterogeneity test is the ***Pseudo F-test***, which compares how much better the solution with k clusters is compared to the solution with $(1- k)$ clusters[22]

$$F_k^2 = \frac{\left(T - \sum_k SKG_k \right) / \left(k - 1 \right)}{\sum_k SKG_k) / \left(n - k \right)}$$

If this statistic has high values, it indicates that $(1- k)$ is a better solution than k solutions.

A *Pseudo t^2-test* is used to compare the means of the pooled clusters for all variables included in the analysis. The statistical significance of separated clusters is determined by the formula:

$$\boldsymbol{t^2 test} = \frac{\left[SKG_t - \left(SKG_r - SKG_s \right) \right] \left(n_R + n_S - 2 \right)}{SKG_r + SKG_s}$$

If the value of the *t^2-test* is statistically significantly higher than the value of the other solutions, this is an indicator that the merged clusters are sufficiently separated (*heterogeneous*) and it is not necessary to merge them.

Cluster profiling

Cluster centroids for all variables included in the analysis are particularly useful in the cluster interpretation phase. Interpretation involves examining salient features for each cluster profile and identifying significant differences between clusters. Cluster solutions that do not have significant variation should be examined once more. It should be examined whether the cluster centroids have similarities with the assumed and expected cluster solutions, based on theory or practical experience. Validation is of particular importance in cluster analysis because clusters descriptively show their structure and additional support is needed to test their relevance. Empirical validation confirms the obtained cluster solution. Therefore, it is necessary to create two sub-samples (*random division of samples*), and then the cluster solutions of these samples are compared in order to obtain the consistency of the obtained clusters and cluster profiles. Validation can also be obtained by examining the differences of variables that are not included in the cluster analysis, and for the analysis of which there is a theoretical or practical reason, in order to better explain the variation and the relationship between the clusters. Profiling the cluster model implies determining the center and varying around the center, as well as determining the statistical difference between clusters. Two clusters may or may not be statistically significantly different either qualitatively or quantitatively. Determining the differences between already formed clusters is the method of Discriminant Analysis.[23]

Paradigm of cluster methods in sports and medicine

The probability of success in sports, the speed of recovery from injuries, the effectiveness of treatment for inherited and acquired diseases depends on a large number of factors. Factors that contribute to a good or bad state are partially known empirically. Scientific knowledge confirms that in addition to the known there are also unknown (error factors). A phenomenon can be defined if the causes of the phenomenon are known. Problems in sports and medicine, ie. achievements in sports and medicine are very complex. Success factors are numerous, very specific and act with different intensity. With the help of scientific methods and empirical knowledge, factors that explain a specific phenomenon are selected. Only on the basis of relevant factors, it is justified to carry out the process of forming clusters with the same or similar characteristics, needs and problems. The need for grouping (*clustering*) is present in all ages and periods of life, starting from kindergartens to nursing homes. Cluster methods are applied in sports guidance in medical practice, diagnosis and application of therapies and treatments in psychiatric treatments. Clusters are effective in the treatment of addiction to alcohol, drugs, gambling, suicide, violence, deviant behavior, etc.

The paradigm of empirical research is the case study in the community.[24] The clinical treatment program was formed on the basis of clusters. The formation of clusters was carried out on the basis of qualitative and quantitative methods that are applied in psychiatry. The goal of the research was an objective perception of the patient's condition, so that the treatment and recovery of people with chronic and serious mental illnesses would be effective. Consistent with the social identity approach and the recovery model, to the extent that people identified as "in recovery," they reported better recovery outcomes (*eg, a sense of purpose*) and reduced psychological distress. Furthermore, recovery identity predicted recovery outcomes more strongly than did psychological distress. Quantitative and qualitative data pointed to collective efficacy (*ie, group empowerment*) as a key mediator of these outcomes. These findings are consistent with the recovery model and speak to the utility of the social identity approach for conceptualizing its effectiveness. Treatment efficiency is higher in "clean" clusters. The clinical implications of clusters and diagnoses are best viewed as complementary systems for describing an individual's needs. Correlations of clusters at admission with primary diagnosis were compared with clusters after hospital discharge. The research provides additional information on the relationship between clusters and diagnosis in the inpatient setting. Clusters and diagnoses are best seen as complementary systems for describing an individual's needs, rather than a 1:1 relationship.

Cluster analysis is important for understanding the heterogeneity of clinical disorders, especially those that challenge common distinctions between physical and psychological etiologies. Cluster analysis methods can refine diagnostic criteria to provide more comprehensive and clinically meaningful profiles within conditions. In IBS this includes consideration of psychological aspects such as anxiety and in the future a broader approach including cognitive and behavioral factors. Cluster analysis also has the potential to improve understanding of different treatment responses in different patient subgroups and to provide more personalized treatment to improve recovery.[25] Big and important decisions made in companies are often accompanied by conflicting opinions and a lack of consensus. The profitability of companies in all sectors, starting from design, services, buying and selling, depends on correct decisions. A study was conducted that showed the pragmatic nature of cluster analysis in making the right decisions. K-Means Clustering was used to determine the different perspectives of different groups of employees; managers, experienced engineers, junior engineers, technical and administrative support staff. The results of the 4-cluster and 5-cluster analysis indicate the need for further study of the dynamics of cluster membership. This, as well as numerous studies, indicate the importance of the application of cluster models in the decision-making phase, i.e. making decisions on capital investment, purchase and sale.[26] Investment strategy and profit from shares, the primary goal is profit, that is, profit with the lowest investment risk. The following Cluster Analysis models were used in the research; (*Average linkage, Centroid and Ward's method*), to determine the preferred method. According to the obtained results, Ward's method proved to be the most appropriate and reliable of all methods. In the example, Ward's method is the only method that gave results that could be analyzed realistically, reasonably, and argumentatively. For the reasons mentioned, further investment strategy is based on the result of Ward's method.[27]

Anthropometric dimensions were measured on a sample of 218 young athletes in the chronological age of 13 to 21 years, functional motor status and efficiency index (IE) were determined. Anthropometric dimensions related to: (*weight, height, sitting height and arm span). Functional motor status was assessed by tests: (sitting and reaching, 1 min sitting, push-ups, hand grip, predicted VO2max, medicine ball throwing, speed at 20 m, vertical jump, standing long jump, balance test standing on one leg-rod test*). Three clusters were obtained that were relevant for individual and team sports. The mentioned factors, i.e. clusters are important in making decisions about the types of sports that would best suit athletes to achieve better results.[28] Identical and similar needs for the formation of clusters are generally encountered in the process of education and upbringing in the teaching of physical education, recreational and professional sports, etc. This work is primarily dedicated to improving the effectiveness of physical education teaching with the application of cluster methods.

The aim and problem of the work

In the title of the work, the goal and problem of the work are implicitly defined. The aim of the work refers to the theoretical description and presentation of various algorithms of cluster methods.[29] The primary problem of the paper is an empirically real, realistic presentation of the possibility of applicability and efficiency of Cluster methods in the function of objective selection.

Research methods

The sample of respondents consists of students (girls and boys) aged 10 and 11. The number of students included in the research was 96. A sample of 7 psycho-motor tests under the name "EuroFit-93" was selected to assess the psycho-motor abilities of the students. In the example, we are talking about "EuroFit" tests; endurance, strength and flexibility. The mentioned tests are generally accepted in many schools in Europe and beyond. EuroFit tests do not require expensive and complex equipment, and can be performed in two to three school hours. The detailed measurement and testing methodology was carried out in accordance with the methodology of Eurofit.[30]

Statistical methods

In accordance with the aim of the work, which is the application of cluster analysis for the purpose of objective selection, the method of „*Non-hierarchical grouping"* was applied. It is about a priori classification of students into clusters based on the results achieved in the psychomotor tests. The Euclidean algorithm of the smallest distances was applied. Before applying the Cluster Analysis, a series of statistical procedures and transformation of the original values to

a common *Z-scale* were performed. The original values in the two tests are shown in units where a lower numerical value represents a better result. *"Hand tappin"* and *"Slalom run 5x10m"*. Inverse transformations were calculated to make the listed values compatible with other values. After the standardization of all seven tests by means of summation, one psycho-motor composite variable called "EuroFit" was created. That composite variable was an objective indicator of the value of each student. Thus, subtle distances are established between students so that no two cases are the same. As the goal of the work was to make ie. to form two homogeneous groups regardless of gender, "*Non-Hierarchical Cluster Analysis*" with two clusters was applied, i.e. two groups.[31] After the establishment of two clusters, two reclusterizations by gender were applied. Finally, „*Canonical Discriminant Analysis"*was applied to determine statistical differences between clusters. In the process of cluster and discriminative analysis, a number of accompanying statistical tests were calculated.

Results and discussion

The common composite variable "EuroFit" was subjected to non-hierarchical clustering using the "Euclidean distances" algorithm. Table 1 shows the final distances between clusters. The statistical significance of the distance between clusters is shown in Table 2. The values of the „Une-factor analysis of variance" were tested using the "F-test" with degrees of freedom (df=N-2) in the example df=94. The F-test value=195.75 points to the conclusion that there is a statistically significant difference between clusters one and two Sig.=.00. The number of students belonging to the first and second clusters is very uniform. In the first cluster there were 49 and in the second 47 students, see the values in Table 3. Such a balanced numerical distribution is good for practical, organizational reasons. The goal of clustering is to create homogeneous groups, with the reason that the effects of the teaching process, i.e. the transformation of students' psycho-motor abilities was faster, therefore more efficient, which is the primary goal of the teaching process. If the difference in the number of students between the clusters is large, then the organization and implementation of the teaching process would be difficult. In the example, the proportion of the number of clusters is 49:47 (Table 3).

Table 1 Distance between final cluster centers

Cluster	1	2
1		0.92
2	0.92	

Table 2 Anova

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
EuroFit	20.93	1	0.11	94	195.75	0

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 3 Number of cases in each cluster

Cluster	1	49	(51)%
	2	47	(49)%
Valid		96	(100)%

Given that the clustering was done on the basis of the achieved results without separating girls from boys. The clustering according to gender is particularly interesting. Table 4 shows the structure by gender. It can be seen that in the first cluster there are 17 girls or 35%, and 32 boys or 65%. In the second cluster, the situation is reversed. As you can see, 33 or 70% of girls belong to the second cluster, and 14 or 30% of boys. It is important to point out that the boys in the first cluster showed better results at the general level. Based on the classification respecting gender, the first cluster would be boys, and the second would be girls. The classification method implies that the students individually recognize which cluster they belong to and the exact hierarchy, which position they occupy in the corresponding cluster. Based on the clustering, the following conclusions can be drawn:

1. If they were to form homogeneous groups based on the total results of the "EuroFit" tests, regardless of gender, the first cluster would have 49 students, of which 17 were girls and 32 were boys. The second cluster would consist of 33 girls and 14 boys. The number of clusters is very even, 49:47, which is good from the point of view of the organization of the teaching process.
2. If you were to form clusters based on gender, there would be 17 girls in the first cluster, and 33 in the second, that is, there would be 32 boys in the first cluster, and 14 in the second. As you can see, the differences in the numbers of one and the other group are enormous. In practice, this would mean a dominant number of students in the first cluster, or a dominant number of girls in the second cluster (Table 4).
3. The decision which variant of clustering to apply depends on the possibilities and standards of the school. Namely, it would be justified and optimal to form clusters by gender and perform additional clustering, i.e. make three homogeneous groups of girls and three homogeneous groups of boys.

Table 4 Clusters by gender

	little girls	boys	Total
Cluster 1	17 (35%)	32 (65)%	49 (100)%
Cluster 2	33 (70%)	14 (30)%	47 (100)%
Toral	50	46	96 (100)%

Results and discussion (girl population)

Table 5 shows the cluster centroid distance in the population of girls. The statistical significance of the distance of the cluster centers is shown in Table 6. Therefore, analysis of variance Anova ie. coefficient F=97.03 shows the statistical significance of differences between clusters which is extremely high Sig.=.00. In the example, the error is less than .01%. It is evident that there are two clusters that are significantly different from each other. The problem is the uneven number of female students in the clusters. As shown in Table 7, the number of female students in the first cluster is 17 and in the second 33. With this procedure, the problem of group homogeneity is solved, but the practical problem related to the implementation and efficiency of the teaching process is not. For this reason, three clusters were formed. The results of the three clusters are shown in Tables 8–10. The first cluster consists of 13 girls who showed better results than the other two. The second cluster is again the largest, consisting of 22 girls, and the third cluster has 15 female students. Table 10 shows the centroid distance between clusters. The greatest distance is between the first and third clusters. The statistical significance of the distance between the clusters is shown in Table 10. The size of the F-coefficient and the probability of error Sig.=.00., indicate a significant difference in terms of quality between the groups. The first group is the smallest but also the best.

Table 5 Distances between final cluster centers

Cluster	1	2
1		0.92
2	0.92	

Table 6 Anova

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
EuroFit	9.49	1	0.098	48	97.03	0

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 7 Number of cases in each cluster

Cluster	1	17 (34)%
	2	33 (66)%
Valid		50 (100)%

Table 8 Distances between final clusters enters

GENDER	Cluster	1	2	3
little girls	1		0.736	1.292
	2	0.736		0.556
	3	1.292	0.556	

Table 9 Anova

GENDER	Cluster		Error		F	Sig.
little giris	Mean Square	df	Mean Square	df		
Mean-Z	5.821	2	0.054	47	107.61	0

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 10 Number of cases in each cluster

little girls	Cluster	1	13	(26)%
		2	22	(44)%
		3	15	(30)%
	Valid		50	(100)%

Based on the obtained results, the following conclusion follows, i.e. recommendation; Special attention should be paid to each cluster, i.e. to the group. Special attention should be paid to girls in clusters one and three. Evidently, the same work program would least benefit the first and third groups (cluster). Assuming the application of the same work program. Female students in the first cluster would progress more slowly, because the existing program is insufficiently demanding, that is, the same program for female students in the third cluster would be too demanding. Objectively, the existing work program would be most suitable for students in the second cluster.

Work, study, training and any transformation in homogeneous and smaller groups is more effective. The general problems of working with several groups are organizational, spatial, financial and personnel. The mentioned factors significantly affect the efficiency of the transformation process.

Results and discussion (boy population)

The distance of the centroid in the population of boys is shown in Table 11. The statistical significance of the distance was determined using the One-Factor Analysis of Variance Anova Table 12. The size of the coefficient F=74.19 was tested for the level of statistical significance. The value Sig.=.00 points to the conclusion that the cluster centroids are statistically significantly different. Thus, there are two groups of boys that differ statistically significantly based on the criteria of motor skills known as EuroFit tests.

Table 11 Distances between final cluster centers

Cluster	1	2
1		0.95
2	0.95	

Table 12 Anova

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
EuroFit	9.13	2	0.12	44	74.19	0

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters.

Table 13 shows the cluster structure in the population of boys. It is easy to notice that there are 31 cases or 67% in the first cluster and 15 or 33% of cases in the second. As you can see, the size (number) of clusters is uneven but at the same time homogeneous. The first cluster consists of students who achieved better results in psychomotor tests. According to the same algorithm, three clusters were formed in the population of boys. The distribution of the three clusters is shown in Table 14.

Table 13 Dječaci number of cases in each cluster

Cluster	1	31 (67%)
	2	15 (33%)
Valid		46 (100%)

Table 14 Number of cases in each cluster

Cluster	1	9 (20%)
	2	22 (48%)
	3	15 (32%)
Valid		46 (100%)

The results in Table 14 indicate that the number of the second cluster dominates, the same as in the population of girls with the difference that in boys the second cluster was obtained by clustering the first, and in girls the second cluster was obtained by clustering the second cluster.

Conclusions based on three clusters Nine boys belong to the highest quality ie. to the first cluster. Twenty-two boys are in the second and fifteen boys are in the third cluster. In this research, clustering was done on the basis of the results achieved in psychomotor tests, so that the first cluster consisted of students with the best results, and the third cluster consisted of cases with the worst results. The recommendations of what should be done in order to improve psychomotor skills are as follows: The teaching process should be organized by gender, as indicated by the results in Table 4. Based on the clusters, Tables 15&16, the teaching process should be implemented with three homogeneous groups separated by gender.

Table 15 Distances between final cluster centers

	Cluster	1	2	3
GENDER	1		0.84	1.933
little girls	2	0.84		1.093
	3	1.933	1.093	

Table 16 Anova

GENDER		Cluster		Error		F	Sig.
little girls		Mean Square	df	Mean Square	df		
boys	EuroFit	5.454	2	0.085	43	64.482	0

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 17 shows relevant basic statistical values (*Mean, Std. Deviaton, Skewess, Kurtosis, Minimum and Maximum*). The total population was 96 girls and boys. The average value of the composite variable of motor skills "EuroFit" *Mean*=8.0 and *Std.Dev*.=.57. Skewness and Kurtosis values as well as *Min*. and *Max*. point to the conclusion that the distribution moves within the limits of normality. The stated statement is visually confirmed by Graph 1. The normality of the distribution was tested using Kolmogorov-Smirnov tests (*K-S*). The value of *K-S, Z*=.485 corresponds to *Asymp.Sig*.=.973 and the value $p<.001$ indicates that there is no statistically significant difference between empirical and theoretical normal distribution.

Table 17 EuroFit test statistics

EuroFit	N=96
Mean	8
Std. Deviation	0.57
Skewness	0.012
Kurtosis	-0.45
Minimum	6.79
Maximum	9.16
Kolmogorov-Smirnov Z	0.485
Asymp. Sig. (2-tailed)	0.973
K-S d=.05	p<.001

In the first step, Non-Hierarchical Cluster Analysis was applied with the application of the Squared Euclidean distance algorithm. Table 18 shows the values of the first cluster. It should be emphasized that the clusters were made according to the achieved results of the composite EuroFit test. Therefore, the clusters are not formed according to gender, but exclusively according to the results achieved in the composite EuroFit test.

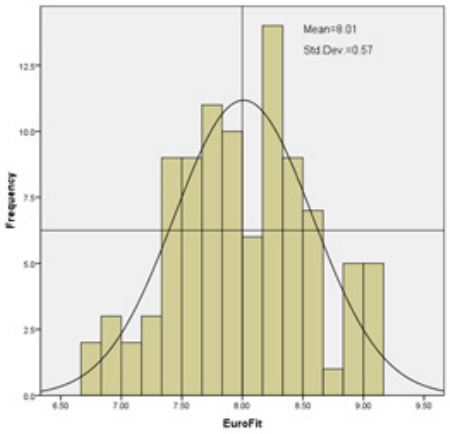
The analogy of the analysis leads us to the second cluster, which unites 47 students of both sexes, table 19. Given that the clusters were formed on the basis of quality, it is expected that the arithmetic mean of the second cluster is lower than the arithmetic mean of the first cluster and the arithmetic mean of the total population (7.55<8.00<8.48). The standard deviation and other measures of variation, including the K-S test values, point to the conclusion that it is a distribution that moves within the limits of normality. Based on the Skewness value, it can be concluded that the first cluster is positively and the second negatively asymmetric. See values in Tables 18 & 19 as well as Graphs 1 and 2.

Table 18 Cluster 1 statistics

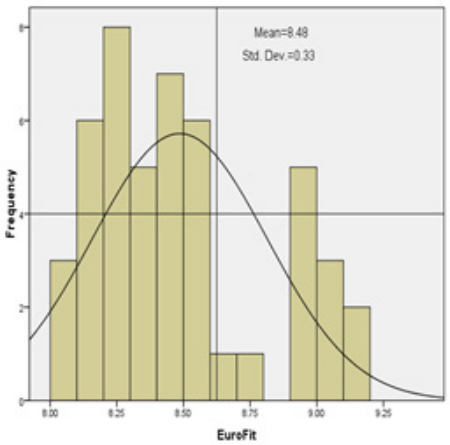
EuroFit	N=49
Mean	8.48
Std. Deviation	0.32
Skewness	0.721
Kurtosis	-0.615
Minimum	8.03
Maximum	9.16
Kolmogorov-Smirnov Z	0.92
Asymp. Sig. (2-tailed)	0.37
K-S d=.05	p<.05

Table 19 Cluster 2 statistics

EirpFit	N=47
Mean	7.55
Std. Deviation	0.32
Skewness	-0.865
Kurtosis	0.014
Minimum	6.79
Maximum	7.99
Kolmogorov-Smirnov Z	1.06
Asymp. Sig. (2-tailed)	0.21
K-S d=.05	p<.01



Graph 1 Cluster 1.



Graph 2 Cluster 2.

In the first cluster there are 49 students of both sexes. Given that it is a hierarchy of quality, the arithmetic mean of the first cluster is expected to be higher than the arithmetic mean of the total population. Chart 1 as well as the values in Table 18 show that the values are within the limits of the normal distribution.

Conclusions based on the results of the cluster analysis

If they were to form homogeneous groups based on the total results of the EuroFit tests regardless of gender, the first cluster would consist of 49 students, of which 17 were girls and 32 were boys. The second cluster would consist of 33 girls and 14 boys. Based on the total number regardless of gender, the size of the clusters is very uniform 49:47 or 51%:49%, which is good from the aspect of the organization of the teaching process, but at the same time bad from the aspect of the efficiency of the EuroFit program.

If the clusters were to be formed on the basis of gender, according to the obtained results, in the first cluster there would be 17 girls or 35%, or 32 boys or 65%. As you can see, the absolute and relative differences between girls and boys are huge. The first cluster is dominated by boys, and the second by girls. From the aspect of greater efficiency of transformation of psycho-motor abilities, the formation of homogeneous groups should be done according to gender.

As it has been proven, the EuroFit program is too demanding for girls, that is, insufficiently stimulating for boys. In order to find optimal solutions, a series of clusterings according to gender and achieved results on psycho-motor tests were performed.

Three clusters in the girls' group and three clusters in the boys' group are recommended. With the difference that in the population of girls, the third cluster was obtained after the clustering of the second cluster, and in the population of boys, the third cluster was obtained after the clustering of the first cluster.

Recommendations on what should be done in order to improve the psycho-motor abilities of girls and boys at this age are as follows: The teaching process should be organized and implemented separately for girls and boys.

Form three homogeneous groups based on the results of the cluster analysis. Implement the teaching process in accordance with the EuroFit program.

Acknowledgments

None.

Conflicts of interest

The author declares that there are no conflicts of interest.

Funding

None.

References

1. Staff Writer. What is cluster analysis? Overview and examples. 2023.
2. Cluster analysis - Wikipedia.
3. Lahav A, Talmon R, Kluger Y. Mahalanobis distance informed by clustering graphic. *A Journal of the IMA*. 2019;8(2):377–406.
4. János Abonyi, Balázs Feil. Cluster analysis for data mining and system identification. 2007.
5. Edwin SD, Camilla LN & Duncan E Statistical power for cluster analysis. *BMC Bioinformatics*. 2022; 23(1):205.
6. Abonyi JS, Feil B. Cluster analysis for data mining and system identification. Boston and Basel, Switzerland: Birkhäuser Basel; 2007.
7. Derrick SB. Determination of the number of clusters in a data set: a stopping rule × clustering algorithm comparison.
8. Hartigan JA. Statistical theory in clustering. *Journal of Classification*. 1985;2:63–76.
9. Euclidean Distance Formula- Derivation, Examples- Cuemath.
10. Joshua David Nelson. On K-means cluste-ring using mahalanobis distance. 2012.
11. The right distance for the clustering. Maybe Mahalanobis?
12. Chaitanya RP. Understanding the concept of Hierarchical clustering Technique. 2018.
13. Tim Bock. What is hierarchical clustering in data analysis?
14. What is the k-nearest neighbors algorithm?
15. Dendrogram by the method of complete linkage -farthest neighbor.
16. Simon Mukwembi. Average distance, minimum degree, and irregularity index. *Discrete Mathematics*. 2024;347(1).
17. Clustering Methods- SAS Help Center.
18. Anna Großwendt. Analysis of Ward's Method. *ArXiv*. 2019.
19. Konni Callista Asysyifaa. Non-hierarchical cluster analysis (K-Means) using R - Medium.2021.
20. Difference between hierarchical and non hierarchical clustering. 2022.
21. Veronika Harantová. Two-step cluster analysis of passenger mobility segmentation. *Mathematics*. 2023.
22. Satoru Hayasaka. How many clusters? Methods for choosing the right number of clusters. 2022.
23. Saleem S. The essence of cluster profiling. 2023.
24. Migliore A, Rossi Lamastra C. Cluster analysis methodological approaches for workplace research and management. 2023;95–107.
25. Windgassen S, Moss-Morris R, Goldsmith et al. The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome. *Journal of Mental Health*. 2017;94–96.
26. Z Aytan Ediz. A model for make-or-buy decisions in engineering design services sector: A case study from Turkey. *International Journal of Innovation and Technology Management*. 2017.
27. Sarah Bjärkbya. Cluster analysis of stocks to define an investment strategy. 2019.
28. Noor Aishah Kamarudin. A cluster analysis of identifying team and individual sports athlete based on anthropometric, health and skill related components. 2022;561–570.
29. Mijanovic M. Effects of the EuroFit program. *MOJ Sports Med*. 2022;5(4):108–114.
30. Statistical methods.
31. Morrison DF. Multivariate statistical methods.